Asymptotic distribution of the distance-based classifier under a strongly spiked eigenvalue model (Bayes Inference and Its Related Topics)

AUTHOR(S):

Ishii, Aki

CITATION:

Ishii, Aki. Asymptotic distribution of the distance-based classifier under a strongly spiked eigenvalue model (Bayes Inference and Its Related Topics). 数理解析研究所講究録 2017, 2047: 1-9

ISSUE DATE:

2017-10

URL:

http://hdl.handle.net/2433/237024

数理解析研究所講究録
第2047巻 2017年 1-9

Asymptotic distribution of the distance-based classifier under a strongly spiked eigenvalue model

東京理科大学・情報科学科　石井 晶 (Aki Ishii)
Department of Information Sciences
Tokyo University of Science

Abstract: We consider two-class classification for high-dimensional data. We consider the distance-based classifier given by Aoshima and Yata (2014). We provide an asymptotic distribution of the classifier under a strongly spiked eigenvalue model.

Key words and phrases: Asymptotic distribution, Distance-based classifier, HDLSS, Large p small n.

1. Introduction

Nowadays, you can see many types of high-dimensional data such as genetic microarrays, medical imaging, text recognition, finance, chemometrics, and so on. A common feature of high-dimensional data is that the data dimension is extremely high, however, the sample size is relatively low. We call such data "HDLSS" or "large p, small n" data, where p is the data dimension and n is the sample size. In this pepar, we consider two-class classification in HDLSS context. We aim to give an asymptotic distribution of the distance-based classifier under a strongly spiked eigenvalue model that was proposed by Aoshima and Yata (2017).

Suppose we have two classes π_i, $i=1,2$, and define independent $p\times n_i$ data matrices, $\boldsymbol{X}_i=[\boldsymbol{x}_{i1},...,\boldsymbol{x}_{in_i}]$, $i=1,2$, from π_i, $i=1,2$, where $\boldsymbol{x}_{ij}$, $j=1,...,n_i$, are independent and identically distributed (i.i.d.) as a p-dimensional distribution with a mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ $(\geq \boldsymbol{O})$. We assume $n_i\geq 3$, $i=1,2$. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ is given by

$$\boldsymbol{\Sigma}_i=\boldsymbol{H}_i\boldsymbol{\Lambda}_i\boldsymbol{H}_i^T=\sum_{s=1}^{p}\lambda_{s(i)}\boldsymbol{h}_{s(i)}\boldsymbol{h}_{s(i)}^T,$$

where $\boldsymbol{\Lambda}_i=\mathrm{diag}(\lambda_{1(i)},...,\lambda_{p(i)})$ having $\lambda_{1(i)}\geq\cdots\geq\lambda_{p(i)}(\geq 0)$ and $\boldsymbol{H}_i=[\boldsymbol{h}_{1(i)},...,\boldsymbol{h}_{p(i)}]$ is an orthogonal matrix of the corresponding eigenvectors. Let $\boldsymbol{X}_i-[\boldsymbol{\mu}_i,...,\boldsymbol{\mu}_i]=\boldsymbol{H}_i\boldsymbol{\Lambda}_i^{1/2}\boldsymbol{Z}_i$ for $i=1,2$. Then, $\boldsymbol{Z}_i$ is a $p\times n_i$ sphered data matrix from a distribution with the zero mean and identity covariance matrix. Let $\boldsymbol{Z}_i=[\boldsymbol{z}_{1(i)},...,\boldsymbol{z}_{p(i)}]^T$ and $\boldsymbol{z}_{j(i)}=(z_{j1(i)},...,z_{jn_i(i)})^T$, $j=1,...,p$, for $i=1,2$. Note that $E(z_{jk(i)}z_{j'k(i)})=0$ $(j\neq j')$ and $\mathrm{Var}(\boldsymbol{z}_{j(i)})=\boldsymbol{I}_{n_i}$, where $\boldsymbol{I}_{n_i}$ denotes the n_i-dimensional identity matrix. Also, note that if $\boldsymbol{X}_i$ is Gaussian, $z_{jk(i)}$s are i.i.d. as the standard normal distribution, $N(0,1)$. We assume that the fourth moments of each variable in $\boldsymbol{Z}_i$ are uniformly bounded for $i=1,2$. Let $\boldsymbol{z}_{oj(i)}=\boldsymbol{z}_{j(i)}-(\bar{z}_{j(i)},...,\bar{z}_{j(i)})^T$, $j=1,...,p$; $i=1,2$, where $\bar{z}_{j(i)}=n_i^{-1}\sum_{k=1}^{n_i}z_{jk(i)}$.

We assume that $P(\lim_{p\to\infty}||\boldsymbol{z}_{o1(i)}|| \neq 0) = 1$ for $i = 1, 2$, where $||\cdot||$ denotes the Euclidean norm.

Let $\boldsymbol{x}_0$ be an observation vector of an individual belonging to π_i $(i = 1, 2)$. We assume $\boldsymbol{x}_0$ and $\boldsymbol{x}_{ij}$s are independent. We estimate $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ by $\overline{\boldsymbol{x}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$ and $\boldsymbol{S}_{in_i} = \sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})^T/(n_i - 1)$. A typical classification rule is that one classifies an individual into π_1 if

$$\begin{aligned}&(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1})^T \boldsymbol{S}_{1n_1}^{-1}(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}) - \log\left\{\frac{\det(\boldsymbol{S}_{2n_2})}{\det(\boldsymbol{S}_{1n_1})}\right\} \\ &< (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2})^T \boldsymbol{S}_{2n_2}^{-1}(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}),\end{aligned} \tag{1.1}$$

and into π_2 otherwise. However, the inverse matrix of $\boldsymbol{S}_{in_i}$ does not exist in the HDLSS context $(p > n_i)$. When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, Bickel and Levina (2004) considered the inverse matrix defined by only diagonal elements of the pooled sample covariance matrix. Yata and Aoshima (2012) considered using a ridge-type inverse covariance matrix derived by the *noise reduction (NR) methodology*. When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, Dudoit et al. (2002) considered using the inverse matrix defined by only diagonal elements of $\boldsymbol{S}_{in_i}$. Aoshima and Yata (2011,2015a) considered substituting $\{\mathrm{tr}(\boldsymbol{S}_{in_i})/p\}\boldsymbol{I}_p$ for $\boldsymbol{S}_{in_i}$ by using the difference of a geometric representation of HDLSS data from each π_i. Aoshima and Yata (2015b) considered quadratic classifiers in general and discussed asymptotic properties and optimality of the classifies under high-dimensional settings. They showed that the misclassification rates tend to zero as the dimension goes to infinity. On the other hand, Hall et al. (2005, 2008), and Chan and Hall (2009) considered distance-based classifiers. Aoshima and Yata (2014) gave the misclassification rate adjusted classifier for multiclass, high-dimensional data whose misclassification rates are no more than specified thresholds under the following condition for eigenvalues:

$$\frac{\lambda_{1(i)}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \to 0 \ \text{ as } p \to \infty \text{ for } i = 1, 2. \tag{1.2}$$

Recently, Aoshima and Yata (2017) considered the "strongly spiked eigenvalue (SSE) model" as follows:

$$\liminf_{p\to\infty}\left\{\frac{\lambda_{1(i)}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}\right\} > 0 \ \text{ for } i = 1 \text{ or } 2. \tag{1.3}$$

On the other hand, Aoshima and Yata (2017) called (1.2) the "non-strongly spiked eigenvalue (NSSE) model".

In this paper, we consider the distance-based classifier under one of the SSE models.

Remark 1.1. For a spiked model such as

$$\lambda_{s(i)} = a_{s(i)}p^{\alpha_{s(i)}} \ (s = 1, ..., t_i) \quad \text{and} \quad \lambda_{s(i)} = c_{s(i)} \ (s = t_i + 1, ..., p) \tag{1.4}$$

with positive and fixed constants, $a_{s(i)}$s, $c_{s(i)}$s and $\alpha_{s(i)}$s, and a positive and fixed integer t_i, note that (1.2) holds when $\alpha_{1(i)} < 1/2$ for $i = 1, 2$. On the other hand, (1.3) holds for the spiked model in (1.4) with $\alpha_{1(i)} \geq 1/2$. See Yata and Aoshima (2012) for the details of the spiked model.

2. Distance-based classifier

Aoshima and Yata (2014) considered a classification rule given by using the identity matrix $\boldsymbol{I}_p$ instead of $\boldsymbol{S}_{in_i}$ in (1.1) as follows: One classifies an individual into π_1 if

$$\Big(\boldsymbol{x}_0 - \frac{\overline{\boldsymbol{x}}_{1n_1} + \overline{\boldsymbol{x}}_{2n_2}}{2}\Big)^T (\overline{\boldsymbol{x}}_{2n_2} - \overline{\boldsymbol{x}}_{1n_1}) - \frac{\mathrm{tr}(\boldsymbol{S}_{1n_1})}{2n_1} + \frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{2n_2} < 0 \tag{2.1}$$

and into π_2 otherwise. Here, $-\mathrm{tr}(\boldsymbol{S}_{1n_1})/(2n_1) + \mathrm{tr}(\boldsymbol{S}_{2n_2})/(2n_2)$ is a bias-correction term. They showed the asymptotic normality of the classifier and provide a sample size determination so as to control misclassification rates being no more than a prespecified value. They further developed the classifier to multiclass classification.

Remark 2.1. Chan and Hall (2009) considered a scale adjusted distance-based classifier as follows: One classifies an individual into π_1 if

$$\sum_{j=1}^{n_1} \frac{||\boldsymbol{x}_0 - \boldsymbol{x}_{1j}||^2}{n_1} - \sum_{j=1}^{n_2} \frac{||\boldsymbol{x}_0 - \boldsymbol{x}_{2j}||^2}{n_2} - \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} \frac{||\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j}||^2}{2n_1(n_1-1)} + \sum_{i=1}^{n_2}\sum_{j=1}^{n_2} \frac{||\boldsymbol{x}_{2i} - \boldsymbol{x}_{2j}||^2}{2n_2(n_2-1)} < 0 \tag{2.2}$$

and into π_2 otherwise. We note that the classifier given by (2.1) is equivalent to the one given by (2.2), though the description of (2.1) is much simpler than (2.2).

We denote the error of misclassifying an individual from π_1 (into π_2) or π_2 (into π_1) by $e(1)$ or $e(2)$, respectively. Let $\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$ and

$$W(\boldsymbol{x}_0) = \Big(\boldsymbol{x}_0 - \frac{\overline{\boldsymbol{x}}_{1n_1} + \overline{\boldsymbol{x}}_{2n_2}}{2}\Big)^T (\overline{\boldsymbol{x}}_{2n_2} - \overline{\boldsymbol{x}}_{1n_1}) - \frac{\mathrm{tr}(\boldsymbol{S}_{1n_1})}{2n_1} + \frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{2n_2}.$$

Aoshima and Yata (2014) considered asymptotic properties of $W(\boldsymbol{x}_0)$ under the following assumptions:

(A-i) $\dfrac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\Delta^2} \to 0$ as $p \to \infty$ for $i = 1, 2$;

(A-ii) $\dfrac{\max_{j=1,2} \mathrm{tr}(\boldsymbol{\Sigma}_j^2)}{n_i \Delta^2} \to 0$ as $p \to \infty$ either when n_i is fixed or $n_i \to \infty$ for $i = 1, 2$.

Then, they gave the asymptotic consistency:

Theorem 2.1 (Aoshima and Yata, 2014). *Assume (A-i) and (A-ii). It holds that as* $p \to \infty$

$$\frac{W(\boldsymbol{x}_0)}{\Delta} = \frac{(-1)^i}{2} + o_p(1) \quad \textit{when } \boldsymbol{x}_0 \in \pi_i$$

for $i = 1, 2$. *Then, the classification rule given by (2.1) has that as* $p \to \infty$

$$e(1) \to 0 \quad \textit{and} \quad e(2) \to 0.$$

Remark 2.2. Under the condition that $\max_{j=1,2}\{\mathrm{tr}(\boldsymbol{\Sigma}_j^2)\}/\Delta^2 \to 0$ as $p \to \infty$, one can claim Theorem 2.1 either when n_i is fixed or $n_i \to \infty$ for $i = 1, 2$.

Aoshima and Yata (2014) also showed the asymptotic normality of (2.1). They assume a general factor model as follows:

$$\boldsymbol{x}_{ij} = \boldsymbol{\Gamma}_i \boldsymbol{w}_{ij} + \boldsymbol{\mu}_i$$

for $j = 1, ..., n_i$; $i = 1, 2$, where $\boldsymbol{\Gamma}_i$ is a $p \times r_i$ matrix for some $r_i > 0$ such that $\boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_i^T = \boldsymbol{\Sigma}_i$, and $\boldsymbol{w}_{ij}$, $j = 1, ..., n_i$, are i.i.d. random vectors having $E(\boldsymbol{w}_{ij}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{w}_{ij}) = \boldsymbol{I}_{r_i}$. As for $\boldsymbol{w}_{ij} = (w_{i1j}, ..., w_{ir_ij})^T$, $i = 1, 2$, we assume that

(A-iii) The fourth moments of each variable in $\boldsymbol{w}_{ij}$ are uniformly bounded, $E(w_{iqj}^2 w_{isj}^2) = 1$ and $E(w_{iqj} w_{isj} w_{itj} w_{iuj}) = 0$ for all $q \neq s, t, u$.

If π_i is $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, (A-iii) naturally follows. Also, Aoshima and Yata (2014) assume the following assumption for $\boldsymbol{\Sigma}_i$, $(i = 1, 2)$.

(A-iv) $\dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_l)}{\mathrm{tr}(\boldsymbol{\Sigma}_j^2)} \in (0, \infty)$ as $p \to \infty$ for $i, j, l = 1, 2$.

Here, $f(p) \in (0, \infty)$ as $p \to \infty$ denotes that $\liminf_{p\to\infty} f(p) > 0$ and $\limsup_{p\to\infty} f(p) < \infty$ for a function $f(\cdot)$. Let

$$\kappa_i = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{n_i} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2)}{n_j} + \sum_{i=1}^{2} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{2n_i(n_i - 1)}$$

for $i (\neq j) = 1, 2$. Let

$$n_{\min} = \min\{n_1, n_2\}.$$

We assume the following assumption:

(A-v) $\dfrac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\kappa_i} = o(1)$ as $p \to \infty$ and $n_{\min} \to \infty$ for $i = 1, 2$.

Then, they have the following result.

Theorem 2.2 (Aoshima and Yata, 2014). *Assume (1.2). Assume also (A-iii) to (A-v). Then, we have that as $p \to \infty$ and $n_{\min} \to \infty$*

$$\frac{W(\boldsymbol{x}_0) - (-1)^i \Delta/2}{\sqrt{\kappa_i}} \Rightarrow N(0, 1) \quad \textit{when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2, \tag{2.3}$$

where "$\Rightarrow$" denotes the convergence in distribution and $N(0, 1)$ denotes a random variable distributed as the standard normal distribution.

Remark 2.3. From Theorem 2.2, for the classification rule by (2.1), it holds that as $p \to \infty$ and $n_{\min} \to \infty$

$$e(i) = \Phi\Big(\frac{-\Delta}{2\sqrt{\kappa_i}}\Big) + o(1) \quad \text{when } \boldsymbol{x}_0 \in \pi_i \text{ for } i = 1, 2 \tag{2.4}$$

under the assumptions in Theorem 2.2, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

3. Asymptotic distribution under a SSE model

In this section, we provide the asymptotic distribution of (2.1) under a SSE model. Now, we consider the following assumptions for each π_i, $i = 1, 2$.

(B-i) $\dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2) - \lambda_{1(i)}^2}{\lambda_{1(i)}^2} = o(1)$ as $p \to \infty$;

(B-ii) $\dfrac{\sum_{r,s\geq 2}^{p} \lambda_{r(i)}\lambda_{s(i)} E\{(z_{rk(i)}^2 - 1)(z_{sk(i)}^2 - 1)\}}{n_i \lambda_{1(i)}^2} = o(1)$ as $p \to \infty$ either when n_i is fixed or $n_i \to \infty$;

(B-iii) $z_{1k(i)}$, $k = 1, ..., n_i$ i.i.d. as $N(0, 1)$.

Note that (B-i) is one of the SSE models. By using the NR method, $\lambda_{j(i)}$s are estimated by

$$\tilde{\lambda}_{j(i)} = \hat{\lambda}_{j(i)} - \frac{\mathrm{tr}(\boldsymbol{S}_{in_i}) - \sum_{s=1}^{j} \hat{\lambda}_{s(i)}}{n_i - 1 - j} \quad (j = 1, ..., n_i - 2), \tag{3.1}$$

where $\hat{\lambda}_{j(i)}$ is the jth sample eigenvalue for $i = 1, 2$. Note that $\tilde{\lambda}_{j(i)} \geq 0$ w.p.1 for $j = 1, ..., n_i - 2$. Yata and Aoshima (2012, 2013, 2016) showed that $\tilde{\lambda}_{j(i)}$ has several consistency properties when $p \to \infty$ and $n_i \to \infty$. On the other hand, when $p \to \infty$ while n_is are fixed, Ishii et al. (2016) gave the following results.

Theorem 3.1 (Ishii et al. 2016). *Under (B-i) and (B-ii), it holds that as $p \to \infty$*

$$\frac{\tilde{\lambda}_{1(i)}}{\lambda_{1(i)}} = \begin{cases} ||\boldsymbol{z}_{o1(i)}||^2/(n_i - 1) + o_p(1) & \textit{when } n_i \textit{ is fixed,} \\ 1 + o_p(1) & \textit{when } n_i \to \infty \end{cases}$$

for $i = 1, 2$. Under (B-i) to (B-iii), it holds that as $p \to \infty$ when n_i is fixed

$$(n_i - 1)\frac{\tilde{\lambda}_{1(i)}}{\lambda_{1(i)}} \Rightarrow \chi_{n_i - 1}^2 \quad \textit{for } i = 1, 2.$$

Now, we consider the following assumptions.

(B-iv) $\dfrac{\lambda_{1(1)}}{\lambda_{1(2)}} = 1 + o(1)$ and $\boldsymbol{h}_{1(1)}^T \boldsymbol{h}_{1(2)} = 1 + o(1)$ as $p \to \infty$.

(B-v) $\dfrac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\lambda_{1(1)}} = o(n_{\min}^{-1})$ as $p \to \infty$ and $n_{\min} \to \infty$ for $i = 1, 2$.

Note that (B-iv) means that the two class share their first eigenspace. One can check the validity of (B-iv) by using a test procedure given by Ishii et al. (2016).

Now, we consider the asymptotic distribution of (2.1) under the SSE model, (B-i). Let $z_{01(i)} = \boldsymbol{h}_{1(i)}^T(\boldsymbol{x}_0 - \boldsymbol{\mu}_i)/\lambda_{1(i)}^{1/2}$ when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2$. Then, we have the following result.

Lemma 3.1. *Assume (B-i), (B-iv) and (B-v). Then, we have that as* $p \to \infty$ *and* $n_{\min} \to \infty$

$$\frac{W(\boldsymbol{x}_0) - (-1)^i\Delta/2}{\lambda_{1(1)}} = z_{01(i)}(\bar{z}_{1(2)} - \bar{z}_{1(1)}) + o_p(n_{\min}^{-1/2})$$

when $\boldsymbol{x}_0 \in \pi_i$ *for* $i = 1, 2$.

From Lemma 3.1 we have the following result.

Theorem 3.2. *Assume (B-i), (B-iii) to (B-v). Then, we have that as* $p \to \infty$ *and* $n_{min} \to \infty$

$$u_n \frac{W(\boldsymbol{x}_0) - (-1)^i\Delta/2}{\lambda_{1(1)}} \Rightarrow U_1 \times U_2$$

when $\boldsymbol{x}_0 \in \pi_i$ *for* $i = 1, 2$,

where $u_n = (n_1^{-1} + n_2^{-1})^{-1/2}$ *and* $U_i s$ *are mutually independent random variables distributed as* $N(0, 1)$.

Remark 3.1. From Theorem 3.2, for the classification rule by (2.1), it holds that as $p \to \infty$ and $n_{\min} \to \infty$

$$e(i) = P\left\{U_1U_2 \le -u_n\frac{\Delta}{2\lambda_{1(1)}}\right\} + o(1) \quad \text{when } \boldsymbol{x}_0 \in \pi_i \text{ for } i = 1, 2 \tag{3.2}$$

under the assumptions in Theorem 3.2. One can estimate Δ by

$$\widehat{\Delta} = ||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2 - \mathrm{tr}(\boldsymbol{S}_{1n_1})/n_1 - \mathrm{tr}(\boldsymbol{S}_{2n_2})/n_2.$$

The estimator was given by Aoshima and Yata (2011). Then, we can estimate (3.2) by $\widehat{\Delta}$ and $\tilde{\lambda}_{1(i)}$.

Appendix

Proof of Lemma 3.1. We assume $\boldsymbol{x}_0 \in \pi_1$ without loss of generality. It holds that

$$||\overline{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i||^2 - \frac{\mathrm{tr}(\boldsymbol{S}_{in_i})}{n_i} = \sum_{s=1}^{p}\sum_{k\neq k'}\lambda_{s(i)}\frac{z_{sk(i)}z_{sk'(i)}}{n_i(n_i-1)}; \tag{A.1}$$

$$(\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^T(\overline{\boldsymbol{x}}_{1n_1} - \boldsymbol{\mu}_1) = \lambda_{1(1)}z_{01(1)}\bar{z}_{1(1)} + \sum_{s=2}^{p}\lambda_{s(1)}z_{01(s)}\bar{z}_{s(1)}, \tag{A.2}$$

where $\boldsymbol{x}_0 - \boldsymbol{\mu}_1 = \boldsymbol{H}_1 \boldsymbol{\Lambda}_1^{1/2} (z_{01(1)}, ..., z_{01(p)})^T$. By using Chebyshev's inequality, for any $\tau > 0$, under (B-i) and (B-iv), we have that as $p \to \infty$ and $n_{\min} \to \infty$

$$E\Big\{\Big|\sum_{s=1}^{p}\sum_{k\neq k'}\frac{\lambda_{s(i)} z_{sk(i)} z_{sk'(i)}}{n_i(n_i-1)}\Big| \geq \tau n_{\min}^{-1/2}\lambda_{1(1)}\Big\} \leq \frac{\sum_{s=1}^{p}\lambda_{s(i)}^2}{\tau^2\lambda_{1(1)}^2(n_i-1)} \to 0;$$

$$E\Big\{\Big|\sum_{s=2}^{p}\lambda_{s(1)} z_{0s(1)}\bar{z}_{s(1)}\Big| \geq \tau n_{\min}^{-1/2}\lambda_{1(1)}\Big\} \leq \frac{\sum_{s=2}^{p}\lambda_{s(1)}^2}{\tau^2\lambda_{1(1)}^2} \to 0,$$

so that from (A.1) and (A.2)

$$||\overline{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i||^2 - \frac{\mathrm{tr}(\boldsymbol{S}_{in_i})}{n_i} = o_p(n_{\min}^{-1/2}\lambda_{1(1)}) \quad \text{for } i = 1, 2;$$

$$(\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^T(\overline{\boldsymbol{x}}_{1n_1} - \boldsymbol{\mu}_1) = \lambda_{1(1)} z_{01(1)}\bar{z}_{1(1)} + o_p(n_{\min}^{-1/2}\lambda_{1(1)}). \tag{A.3}$$

Let $\beta_{st} = (\lambda_{s(1)}\lambda_{t(2)})^{1/2}\boldsymbol{h}_{s(1)}^T\boldsymbol{h}_{t(2)}$ for all s, t. Then, we write that

$$\begin{aligned}(\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^T(\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2) &= \sum_{s,t\geq 1}^{p}\beta_{st} z_{0s(1)}\bar{z}_{t(2)} \\ &= \beta_{11} z_{01(1)}\bar{z}_{1(2)} + \sum_{s=2}^{p}\beta_{s1} z_{0s(1)}\bar{z}_{1(2)} + \sum_{t=2}^{p}\beta_{1t} z_{01(1)}\bar{z}_{t(2)} \\ &\quad + \sum_{s,t\geq 2}^{p}\beta_{st} z_{0s(1)}\bar{z}_{t(2)}. \end{aligned} \tag{A.4}$$

Let $\boldsymbol{\Sigma}_{i*} = \sum_{s=2}^{p}\lambda_{s(i)}\boldsymbol{h}_{s(i)}\boldsymbol{h}_{s(i)}^T$ for $i = 1, 2$. Here, we have that

$$E\Big\{\Big(\sum_{s=2}^{p}\beta_{s1} z_{0s(1)}\bar{z}_{1(2)}\Big)^2\Big\} = \frac{\sum_{s=2}^{p}\beta_{s1}^2}{n_2} = \frac{\lambda_{1(2)}\boldsymbol{h}_{1(2)}^T\boldsymbol{\Sigma}_{1*}\boldsymbol{h}_{1(2)}}{n_2} \leq \frac{\lambda_{1(2)}\lambda_{2(1)}}{n_2};$$

$$E\Big\{\Big(\sum_{t=2}^{p}\beta_{1t} z_{01(1)}\bar{z}_{t(2)}\Big)^2\Big\} = \frac{\lambda_{1(1)}\boldsymbol{h}_{1(1)}^T\boldsymbol{\Sigma}_{2*}\boldsymbol{h}_{1(1)}}{n_2} \leq \frac{\lambda_{1(1)}\lambda_{2(2)}}{n_2};$$

$$E\Big\{\Big(\sum_{s,t\geq 2}^{p}\beta_{st} z_{0s(1)}\bar{z}_{t(2)}\Big)^2\Big\} = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{1*}\boldsymbol{\Sigma}_{2*})}{n_2} \leq \frac{\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_{1*}^2)\mathrm{tr}(\boldsymbol{\Sigma}_{2*}^2)}}{n_2}.$$

Then, by using Chebyshev's inequality, for any $\tau > 0$, under (B-i) and (B-iv), we have that as $p \to \infty$ and $n_{\min} \to \infty$

$$P\Big(\Big|\sum_{s=2}^{p}\beta_{s1} z_{0s}\bar{z}_{1(2)}\Big| > \tau n_{\min}^{-1/2}\lambda_{1(1)}\Big) \leq \frac{\lambda_{1(2)}\lambda_{2(1)}}{\tau^2\lambda_{1(1)}^2} \to 0;$$

$$P\Big(\Big|\sum_{t=2}^{p}\beta_{1t} z_{01}\bar{z}_{t(2)}\Big| > \tau n_{\min}^{-1/2}\lambda_{1(1)}\Big) \leq \frac{\lambda_{1(1)}\lambda_{2(2)}}{\tau^2\lambda_{1(1)}^2} \to 0;$$

$$P\Big(\Big|\sum_{s,t\geq 2}^{p}\beta_{st} z_{0s}\bar{z}_{t(2)}\Big| > \tau n_{\min}^{-1/2}\lambda_{1(1)}\Big) \leq \frac{\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_{1*}^2)\mathrm{tr}(\boldsymbol{\Sigma}_{2*}^2)}}{\tau^2\lambda_{1(1)}^2} \to 0.$$

Then, from (A.4) we have that

$$(\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^T(\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2) = \beta_{11} z_{01} \bar{z}_{1(2)} + o_p(\lambda_{1(1)} n_{\min}^{-1/2})$$
$$= \lambda_{1(1)} z_{01} \bar{z}_{1(2)} + o_p(\lambda_{1(1)} n_{\min}^{-1/2}). \tag{A.5}$$

Also, under (B-iv) and (B-v), we have that as $p \to \infty$ and $n_{\min} \to \infty$

$$P\Big(\Big|(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{x}_0 - \boldsymbol{\mu}_1)\Big| > \tau n_{\min}^{-1/2} \lambda_{1(1)}\Big) \le \frac{n_{\min}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_1 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\tau^2 \lambda_{1(1)}^2} \to 0;$$
$$P\Big(\Big|(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2)\Big| > \tau n_{\min}^{-1/2} \lambda_{1(1)}\Big) \le \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\tau^2 \lambda_{1(1)}^2} \to 0,$$

so that

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{x}_0 - \boldsymbol{\mu}_1) = o_p(n_{\min}^{-1/2} \lambda_{1(1)});$$
$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2) = o_p(n_{\min}^{-1/2} \lambda_{1(1)}) \tag{A.6}$$

Note that

$$W(\boldsymbol{x}_0) + \Delta/2 = \frac{1}{2} \sum_{i=1}^{2} (-1)^{i+1} \Big\{ \|\overline{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i\|^2 - \frac{\mathrm{tr}(\boldsymbol{S}_{in_i})}{n_i} \Big\}$$
$$+ \sum_{i=1}^{2} (-1)^i (\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^T (\overline{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i)$$
$$- (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T (\boldsymbol{x}_0 - \boldsymbol{\mu}_1) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T (\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2). \tag{A.7}$$

Then, by combining (A.3), (A.5) and (A.6) with (A.7), under (B-i), (B-iv) and (B-v), it holds that as $p \to \infty$ and $n_{\min} \to \infty$

$$\frac{W(\boldsymbol{x}_0) + \Delta/2}{\lambda_{1(1)}} = z_{01(1)}(\bar{z}_{1(2)} - \bar{z}_{1(1)}) + o_p(n_{\min}^{-1/2}).$$

For the case when $\boldsymbol{x}_0 \in \pi_2$, we have the result similarly. Thus the proof is completed.

Proof of Theorem 3.2. By using Lemma 3.1, the result is obtained straightforwardly.

Acknowledgment

I would like to express my sincere gratitude to Professor Makoto Aoshima, for his enthusiastic guidance and helpful support to my research project. I would also like to thank Associate Professor, Kazuyoshi Yata, for his valuable suggestions.

References

Aoshima, M., Yata, K. (2011). Two-stage procedures for high-dimensional data. *Sequential Analysis (Editor's special invited paper), 30*, 356–399.

Aoshima, M., Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics, 66*, 983–1010.

Aoshima, M., Yata, K. (2015a). Geometric classifier for multiclass, high-dimensional data. *Sequential Analysis, 34*, 279–294.

Aoshima, M., Yata, K. (2015b). High-dimensional quadratic classifiers in non-sparse settings. *arXiv preprint*, arXiv:1503.04549.

Aoshima, M., Yata, K. (2017). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica*, in press (arXiv:1602.02491).

Bickel, P.J., Levina, E. (2004). Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. *Bernoulli, 10*, 989–1010.

Chan, Y.-B., Hall, P. (2009). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika, 96*, 469–478.

Dudoit, S., Fridlyand, J., Speed, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97*, 77–87.

Hall, P., Marron, J.S., Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society, Series B, 67*, 427–444.

Hall, P., Pittelkow, Y., Ghosh, M. (2008). Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *Journal of the Royal Statistical Society, Series B, 70*, 159–173.

Ishii, A., Yata, K., Aoshima, M. (2016). Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context. *Journal of Statistical Planning and Inference, 170*, 186–199.

Yata, K., Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis, 105*, 193–215.

Yata, K., Aoshima, M. (2013). PCA consistency for the power spiked model in high-dimensional settings. *Journal of Multivariate Analysis, 122*, 334–354.

Yata, K., Aoshima, M. (2016). Reconstruction of a high-dimensional low-rank matrix. *Electronic Journal of Statistics, 10*, 895–917.